

On Consistency of Graph-based Semi-supervised Learning

Chengan Du and Yunpeng Zhao
 Department of Statistics
 George Mason University

Mar 17, 2017

Abstract

Graph-based semi-supervised learning is one of the most popular methods in machine learning. Some of its theoretical properties such as bounds for the generalization error and the convergence of the graph Laplacian regularizer have been studied in computer science and statistics literatures. However, a fundamental statistical property, the consistency of the estimator from this method has not been proved. In this article, we study the consistency problem under a non-parametric framework. We prove the consistency of graph-based learning in the case that the estimated scores are enforced to be equal to the observed responses for the labeled data. The sample sizes of both labeled and unlabeled data are allowed to grow in this result. When the estimated scores are not required to be equal to the observed responses, a tuning parameter is used to balance the loss function and the graph Laplacian regularizer. We give a counterexample demonstrating that the estimator for this case can be inconsistent. The theoretical findings are supported by numerical studies.

1 Introduction

Semi-supervised learning is a class of machine learning methods that stand in the middle ground between supervised learning in which all training data are labeled, and unsupervised learning in which no training data are labeled. Specifically, in addition to the labeled training data X_1, \dots, X_n , there exist unlabeled inputs X_{n+1}, \dots, X_{n+m} . Under certain assumptions on the geometric structure of the input data, such as the cluster assumption or the low-dimensional manifold assumption [7], the use of both labeled and unlabeled data can achieve better prediction accuracy than supervised learning that only uses labeled inputs X_1, \dots, X_n .

Semi-supervised learning has become popular since the acquisition of unlabeled data is relatively inexpensive. A large number of methods were developed under the framework of semi-supervised learning. For example, [18] proposed that the combination of labeled and unlabeled data will improve the prediction accuracy under the assumption of mixture models. The self-training method [19] and the co-training method [13] were soon applied to semi-supervised learning when mixture models are not assumed. [23] described an approach to semi-supervised clustering based on hidden

Markov random fields (HMRFs) that can combine multiple approaches in a unified probabilistic framework. [1] proposed a probabilistic framework for semi-supervised learning incorporating a K-means type hard partition clustering algorithm (HMRF-Kmeans). [20] proposed the transductive support vector machines (TSVMs) that used the idea of transductive learning by including unlabeled data in the computation of the margin. Transductive learning is a variant of semi-supervised learning which focuses on the inference of the correct labels for the given unlabeled data other than the inference of the general rule. [4] used a convex relaxation of the optimization problem called semi-definite programming as a different approaches to the TSVMs.

In this article, we focus on a particular semi-supervised method – graph-based semi-supervised learning. In this method, the geometric structure of the input data are represented by a graph $G = (V, E)$, where nodes $V = \{v_1, \dots, v_{n+m}\}$ represent the inputs and edges E represent the similarities between them. The similarities are given in an $n + m$ by $n + m$ symmetric similarity matrix (or called *kernel* matrix), $W = [w_{ij}]$, where $0 \leq w_{ij} \leq 1$. The larger w_{ij} implies that X_i and X_j are more similar. Further, let Y_1, \dots, Y_n be the responses of the labeled data.

[25] proposed the following graph-based learning method,

$$\min_{\mathbf{f}=(f_1, \dots, f_{n+m})^T} \sum_{i=1}^{n+m} \sum_{j=1}^{n+m} w_{ij} (f_i - f_j)^2 \quad (1)$$

$$\text{subject to } f_i = Y_i, i = 1, \dots, n.$$

Its solution is called the estimated scores. The objective function (1) (named “hard criterion” thereafter), requires all the estimated score to be exactly the same as the responses for the labeled data. [8] relaxed this requirement by proposing a soft version (named “soft criterion” thereafter). We follow an equivalent form given in [26],

$$\min_{\mathbf{f}=(f_1, \dots, f_{n+m})^T} \sum_{i=1}^n (Y_i - f_i)^2 + \frac{\lambda}{2} \sum_{i=1}^{n+m} \sum_{j=1}^{n+m} w_{ij} (f_i - f_j)^2. \quad (2)$$

The soft criterion belongs to the “loss+penalty” paradigm: It searches for the minimizer $\hat{\mathbf{f}}$ which achieves a small training error, and in the meanwhile imposes the smoothness on $\hat{\mathbf{f}}$ by a penalty based on similarity matrix. It can be easily seen that when $\lambda = 0$ the soft criterion is equivalent to the hard criterion.

Remark The tuning parameter λ being 0 in the soft criterion (2) is understood in the following sense: The squared loss has infinite weight and thereby $Y_i = f_i$ for all labeled data. But $\sum_{i=1}^{n+m} \sum_{j=1}^{n+m} w_{ij} (f_i - f_j)^2$ still plays a crucial role when it has no conflict with the hard constraints on the labeled data, that is, it provides links between f_i ’s on the labeled and unlabeled data. Therefore, the soft criterion (2) at $\lambda = 0$ becomes the hard criterion (1).

Researchers have proposed different variants of graph-based learning methods, such as [24] and [2]. We only focus on (1) and (2) in this article.

The theoretical properties of graph-based learning have been studied in computer science and statistics literatures. [5] derived the limit of the Laplacian regularizer when the sample size of unlabeled data goes to infinity. [11] considered the convergence of Laplacian regularizer on Riemannian manifolds. [3] reinterpreted the graph Laplacian as a measure of intrinsic distances between inputs on a manifold and reformulated the problem as a functional optimization in a reproducing kernel Hilbert space. [16] pointed out that the hard criterion can yield completely noninformative solution when the size of unlabeled data goes to infinity and labeled data are finite, that is, the solution can give a perfect fit on the labeled data but remains as 0 on the unlabeled data. [14] obtained the asymptotic mean squared error of a different version of graph-based learning criterion. [2] gave a bound of the generalization error for a slightly different version of (2).

But to the best of our knowledge, no result is available in literature on a very fundamental question – the consistency of graph-based learning, which is the main focus of this article. Specifically, we want to answer the question that under what conditions \hat{f}_i will converge to $\mathbb{E}[Y_i|X_i]$ on unlabeled data, where $\mathbb{E}[Y_i|X_i]$ is the true probability of a positive label given X_i if responses are binary, and $\mathbb{E}[Y_i|X_i]$ is the regression function on X_i if responses are continuous. We will always call $\mathbb{E}[Y_i|X_i]$ as regression function for simplicity.

Most of the literatures discussed above considered a “functional version” of (1) and (2). They used a functional optimization problem with the optimizer $\hat{f}(x)$ being a function, as an approximation of the original problem with the optimizer $\hat{\mathbf{f}}$ being a vector. And they studied the behavior of the limit of graph Laplacian and the solution $\hat{f}(x)$. We do not adopt this framework but use a more direct approach. We focus on the original problem and study the relations of \hat{f}_i and $\mathbb{E}[Y_i|X_i]$ under the general non-parametric setting. Our approach essentially belongs to the framework of transductive learning, which focuses on the prediction on the given unlabeled data X_{n+1}, \dots, X_{n+m} , not the general mapping from inputs to responses. By establishing a link between the optimizer of (1) and the Nadaraya-Watson estimator [15, 21] for kernel regression, we will prove the consistency of the hard criterion. The theorem allows both m and n to grow. On the other hand, we show that the soft criterion is inconsistent for sufficiently large λ . To the best of our knowledge, this is the first result that explicitly distinguishes the hard criterion and the soft criterion of graph-based learning from a theoretical perspective and shows that they have very different asymptotic behaviors.

The rest of the article is organized as follows. In Section 2, we state the consistency result for the hard criterion and give the counterexample for the soft criterion. We prove the consistency result in Section 3. Numerical studies in Section 4 support our theoretical findings. Section 5 concludes with a summary and discussion of future research directions.

2 Main Results

We begin with basic notation and setup. Let $(X_1, Y_1), \dots, (X_{n+m}, Y_{n+m})$ be independently and identically distributed pairs. Here each X_i is a d -dimensional vector and $\mathbf{Y} = (Y_1, \dots, Y_{n+m})^T$ are binary responses labeled as 1 and 0 (the classification case) or continuous responses (the regression case). The last m responses are unobserved.

[25] used a fixed point algorithm to solve the hard criterion (1), which is

$$f_a = \frac{\sum_{i=1}^{n+m} w_{ak} f_i}{\sum_{i=1}^{n+m} w_{ai}}, \quad a = n+1, \dots, n+m. \quad (3)$$

Note that (3) is not a closed-form solution but an updating formula for the iterative algorithm, since its right-hand side depends on unknown quantities.

In order to obtain a closed-form solution for (1), we begin by solving the soft version (2) and then let $\lambda = 0$. Recall that W is the similarity matrix. Let $D = \text{diag}(d_1, \dots, d_{n+m})$ where $d_i = \sum_{j=1}^{n+m} w_{ij}$, and $L = D - W$ being the unnormalized graph Laplacian (see [17] for more details). Soft criterion (2) can be written in matrix form

$$\min_{\mathbf{f}} \sum_{i=1}^n (\mathbf{f} - \mathbf{Y}_n)^T (\mathbf{f} - \mathbf{Y}_n) + \lambda \mathbf{f}^T L \mathbf{f}, \quad (4)$$

where $\mathbf{Y}_n = (Y_1, \dots, Y_n)^T$. Further, let V be an $n+m$ by $n+m$ matrix defined as

$$V = \begin{pmatrix} I_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Then by taking the derivative of (4) with respect to \mathbf{f} and setting equal to zero, we obtain the solution as follows,

$$\hat{\mathbf{f}} = (V + \lambda L)^{-1} V \begin{pmatrix} \mathbf{Y}_n \\ \mathbf{0} \end{pmatrix}.$$

What we are interested in are the estimated scores on the unlabeled data, i.e. $\hat{\mathbf{f}}_{(n+1):(n+m)} = (\hat{f}_{n+1}, \dots, \hat{f}_{n+m})^T$. In order to obtain an explicit form for $\hat{\mathbf{f}}_{(n+1):(n+m)}$, we use a formula for inverse of a block matrix (see standard textbooks on matrix algebra such as [12] for more details): For any non-singular square matrix A

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

$$A^{-1} = \begin{pmatrix} (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & -(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} \\ -(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{pmatrix}.$$

Write D and W as 2×2 block matrices,

$$D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}, W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}.$$

By the formula above,

$$\begin{aligned} \hat{\mathbf{f}}_{(n+1):(n+m)} = \\ (D_{22} - W_{22} - \lambda W_{21}(I_n + \lambda D_{11} - \lambda W_{11})^{-1}W_{12})^{-1}W_{21}(I_n + \lambda D_{11} - \lambda W_{11})^{-1}\mathbf{Y}_n. \end{aligned} \quad (5)$$

Letting $\lambda = 0$, we obtain the solution for the hard criterion (1),

$$\hat{\mathbf{f}}_{(n+1):(n+m)} = (D_{22} - W_{22})^{-1} W_{21} \mathbf{Y}_n. \quad (6)$$

[2] obtained a similar formula for a slightly different objective function.

Clearly, the form of (6) is closely related to the Nadaraya-Watson estimator [15],[21] for kernel regression, which is

$$\hat{q}_{n+a} = \frac{\sum_{i=1}^n w_{n+a,i} Y_i}{\sum_{k=1}^n w_{n+a,i}}, \quad a = 1, \dots, m. \quad (7)$$

The Nadaraya-Watson estimator is well studied under the non-parametric framework. We can construct W by a kernel function, that is, let $w_{ij} = K((X_i - X_j)/h_n)$, where K is a nonnegative function on \mathbb{R}^d , and h_n is a positive constant controlling the bandwidth of the kernel. Let $q(X) = \mathbb{E}[Y|X]$ be the true regression function. The consistency of Nadaraya-Watson estimator was first proved by [21] and [15]. And many other researchers such as [9] and [6] studied its asymptotic properties under different assumptions. Here we follow the result in [10]. If $h_n \rightarrow 0$, $nh_n^d \rightarrow \infty$ as $n \rightarrow \infty$, and K satisfies:

- (i) K is bounded by $k^* < \infty$;
- (ii) The support of K is compact;
- (iii) $K \geq \beta I_B$ for some $\beta > 0$ and some closed ball B centered at the origin and having positive radius δ ,

then \hat{q}_{n+a} converges to $q(X_{n+a})$ in probability for $a = 1, \dots, m$.

By establishing a connection between the solution of the hard criterion and Nadaraya-Watson estimator, we prove the following main theorem:

Theorem 1. *Suppose that $(X_1, Y_1), (X_2, Y_2), \dots, (X_{n+m}, Y_{n+m})$ are independently and identically distributed with Y_i being bounded; h_n and K satisfy the above conditions. Further, let Z be the difference of two independent X 's, i.e. $Z = X_1 - X_2$, with probability density function $g_Z(z)$. Assume that there exists $c > 0$, such that for any $\Delta \leq c$,*

$$\min_{\|z\| \leq \Delta} g_Z(z) = e > 0. \quad (8)$$

Then, for $m = o(nh_n^d)$, \hat{f}_{n+a} given in (5) converges to $q(X_{n+a})$ in probability, for $a = 1, \dots, m$.

The proof will be given in Section 3.

Remark Theorem 1 established the consistency of the hard criterion under the standard non-parametric framework with two additional assumptions. Firstly, both labeled data and unlabeled data are allowed to grow but the size of unlabeled data m grows slower than the size of labeled data n . We conjecture that when m grows faster than n , the graph-based semi-supervised learning

may not be consistent based on the simulation studies in Section 4. [16] also suggested that the method may not work when m grows too fast. Secondly, we assume that density function of the difference of two independent inputs is strictly positive near the origin, which is a mild technical condition valid for commonly used density functions.

Theorem 1 provides some surprising insights about the hard criterion of graph-based learning. At a first glance, the hard criterion makes an impractical assumption that requires the responses to be noiseless, while the soft criterion seems to be a more natural choice. But according to our theoretical analysis, the hard criterion is consistent under the standard non-parametric framework where the responses on training data are of course allowed to be random and noisy.

We now consider the soft criterion with $\lambda \neq 0$.

Proposition 2. *Suppose that $(X_1, Y_1), (X_2, Y_2), \dots, (X_{n+m}, Y_{n+m})$ are independently and identically distributed with $|\mathbb{E}(Y_1)| < \infty$. Further, suppose that W represents a connected graph. Then for sufficiently large λ , the soft criterion (2) is inconsistent.*

Proof. Consider another extreme case of the soft criterion (2), $\lambda = \infty$. When W represents a connected graph, the objective function becomes

$$\min_{\mathbf{f}=(f_1, \dots, f_n)^T} \sum_{i=1}^n (Y_i - f_i)^2 \quad (9)$$

$$\text{subject to } f_i = f_j, 1 \leq i, j \leq n + m.$$

It is easy to check that the solution of (9), denoted by $\hat{\mathbf{f}}(\infty)$, is given by

$$\hat{f}_{n+a}(\infty) = \frac{1}{n} \sum_{i=1}^n Y_i, \quad a = 1, \dots, m.$$

By the law of large numbers,

$$\lim_{n \rightarrow \infty} \hat{f}_{n+a}(\infty) = \mathbb{E}[q(X_1)] \text{ almost surely.}$$

Clearly, $\mathbb{E}[q(X_1)] \neq q(X_{n+a})$ since the right-hand side is a random variable. This implies that for sufficiently large λ , the soft criterion is inconsistent. \square

3 Proof of the Main Theorem

We give the proof of Theorem 1 in this section.

Recall that

$$\hat{\mathbf{f}}_{(n+1):(n+m)} = (D_{22} - W_{22})^{-1} W_{21} \mathbf{Y}_n.$$

We first focus on $(D_{22} - W_{22})^{-1}$. Clearly,

$$(D_{22} - W_{22})^{-1} = (I_m - D_{22}^{-1}W_{22})^{-1}D_{22}^{-1}.$$

For any positive integer l , define

$$S_l = D_{22}^{-1}W_{22} + (D_{22}^{-1}W_{22})^2 + (D_{22}^{-1}W_{22})^3 + \cdots + (D_{22}^{-1}W_{22})^l.$$

Our goal is to prove that the limit of S_l exists with probability approaching 1, and thus we can have

$$(I_m - D_{22}^{-1}W_{22})^{-1} = I_m + \lim_{l \rightarrow \infty} S_l$$

with probability approaching 1 [22].

By definition,

$$D_{22} = \begin{pmatrix} d_{n+1,n+1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_{n+m,n+m} \end{pmatrix}, \quad W_{22} = \begin{pmatrix} w_{n+1,n+1} & \cdots & w_{n+1,n+m} \\ \vdots & \ddots & \vdots \\ w_{n+m,n+1} & \cdots & w_{n+m,n+m} \end{pmatrix},$$

where

$$d_{n+a,n+a} = \sum_{k=1}^{n+m} w_{n+a,k}, \quad w_{n+a,i} = K \left(\frac{X_{n+a} - X_i}{h_n} \right),$$

for $1 \leq a \leq m, 1 \leq i \leq n+m$. Thus we have

$$D_{22}^{-1}W_{22} = \begin{pmatrix} w_{n+1,n+1}/d_{n+1,n+1} & \cdots & w_{n+1,n+m}/d_{n+1,n+1} \\ \vdots & \ddots & \vdots \\ w_{n+m,n+1}/d_{n+m,n+m} & \cdots & w_{n+m,n+m}/d_{n+m,n+m} \end{pmatrix}.$$

Since $h_n \rightarrow 0$, there exist $n_0 \in \mathbb{N}$, such that $\delta h_n \leq c$ holds for every $n > n_0$. Thus by the assumption in (8), for $1 \leq a \leq m$ and $1 \leq i \leq n$,

$$\begin{aligned} p_n &\triangleq \mathbb{E}(I\{\|X_i - X_{n+a}\| \leq \delta h_n\}) \\ &= \mathbb{P}(\|Z\| \leq \delta h_n) = \int_{\|z\| \leq \delta h_n} g_Z(z) \mu dz \geq eV_d(\delta h_n) = sh_n^d, \end{aligned}$$

where $V_d(\delta h_n)$ denotes the volume of a d -dimensional ball with radius h_n , and s is a constant independent with n . Since $nh_n^d \rightarrow \infty$, the above inequality implies $np_n \rightarrow \infty$. On the other side, $p_n \rightarrow 0$ since $h_n \rightarrow 0$.

Further,

$$\text{Var}(I\{\|X_i - X_{n+a}\| \leq \delta h_n\}) = p_n(1 - p_n).$$

By Chebyshev's Inequality, for any $0 < \epsilon < 1/2$, since $nh_n^d \rightarrow \infty$,

$$\begin{aligned}
& \mathbb{P} \left(\left| \frac{\sum_{i=1}^n I\{\|X_i - X_{n+a}\| \leq \delta h_n\}}{np_n} - 1 \right| \geq \epsilon \right) \\
&= \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n I\{\|X_i - X_{n+a}\| \leq \delta h_n\} - p_n \right| \geq \epsilon p_n \right) \\
&\leq \frac{p_n(1-p_n)}{n\epsilon^2 p_n^2} \leq \frac{1}{\epsilon^2 p_n n} \leq \frac{1}{\epsilon^2 s n h_n^d} \rightarrow 0 \quad \text{as } n \rightarrow \infty.
\end{aligned} \tag{10}$$

This further implies

$$\frac{\sum_{i=1}^n I\{\|X_i - X_{n+a}\| \leq \delta h_n\}}{np_n} \rightarrow 1 \quad \text{in probability.}$$

We now continue to study the property of $D_{22}^{-1}W_{22}$. Consider each element $(D_{22}^{-1}W_{22})_{ab}$ of this matrix. For $1 \leq a, b \leq m$,

$$\begin{aligned}
(D_{22}^{-1}W_{22})_{ab} &= \frac{w_{n+a,n+b}}{d_{n+a,n+a}} = K \left(\frac{X_{n+a} - X_{n+b}}{h_n} \right) / \sum_{i=1}^{n+m} K \left(\frac{X_i - X_{n+a}}{h_n} \right) \\
&\leq \frac{k^*}{\beta \sum_{i=1}^n I\{\|X_i - X_{n+a}\| \leq \delta h_n\}},
\end{aligned}$$

by condition (i) and (iii). For simplicity of notation, let

$$\Phi_n(a) = \frac{\sum_{i=1}^n I\{\|X_i - X_{n+a}\| \leq \delta h_n\}}{np_n},$$

where Φ_n is a nonnegative function depending on n . Thus $\frac{k^*}{\beta \Phi_n(a) np_n}$ is an upper bound of every element in the matrix $D_{22}^{-1}W_{22}$. By (10), we have

$$\mathbb{P}(0 \leq \Phi_n(a) \leq 1 - \epsilon) \leq \mathbb{P}(|\Phi_n(a) - 1| \geq \epsilon) \leq \frac{1}{\epsilon^2 s n h_n^d},$$

which implies

$$\begin{aligned}
\mathbb{P} \left(\min_{1 \leq a \leq m} \Phi_n(a) \leq 1 - \epsilon \right) &= \mathbb{P} \left(\bigcup_{a=1}^m \{\Phi_n(a) \leq 1 - \epsilon\} \right) \\
&\leq \sum_{a=1}^m \mathbb{P}(\Phi_n(a) \leq 1 - \epsilon) \leq \frac{m}{\epsilon^2 s n h_n^d},
\end{aligned}$$

and

$$\mathbb{P} \left(\max_{1 \leq a \leq m} \frac{k^*}{\beta \Phi_n(a) np_n} \leq \frac{k^*}{\beta(1 - \epsilon) np_n} \right) \geq 1 - \frac{m}{\epsilon^2 s n h_n^d}.$$

Since $\frac{m}{\epsilon^2 s n h_n^d} \rightarrow 0$, we have

$$\mathbb{P} \left(\max_{1 \leq a, b \leq m} (D_{22}^{-1} W_{22})_{ab} \leq \max_{1 \leq a \leq m} \frac{k^*}{\beta \Phi_n(a) n p_n} \leq M \frac{1}{n h_n^d} \right) \rightarrow 1, \quad \text{as } n \rightarrow \infty, \quad (11)$$

where $M = \frac{2k}{s\beta} > \frac{k^*}{(1-\epsilon)s\beta}$. Note that M is a constant independent with n and m .

For the sake of simplicity, we say a matrix A has *tiny elements*, if

$$\|A\|_{\max} \leq M \frac{1}{n h_n^d},$$

with probability approaching 1, where $\|A\|_{\max} = \max_{ij} A_{ij}$. And $(A)_i$ denotes the i -th row of A . Then $D_{22}^{-1} W_{22}$ has tiny elements by (11). Moreover,

$$\begin{aligned} \|(D_{22}^{-1} W_{22})^2\|_{\max} &= \|(D_{22}^{-1} W_{22})(D_{22}^{-1} W_{22})\|_{\max} \\ &\leq (M \frac{1}{n h_n^d})^2 m = \frac{M}{n h_n^d} (\frac{mM}{n h_n^d}) \end{aligned}$$

holds with probability approaching 1. By induction,

$$\|(D_{22}^{-1} W_{22})^l\|_{\max} = \|(D_{22}^{-1} W_{22})(D_{22}^{-1} W_{22})^{l-1}\|_{\max} \leq \frac{M}{n h_n^d} (\frac{mM}{n h_n^d})^{l-1},$$

with probability approaching 1. Therefore,

$$\begin{aligned} \|S_l\|_{\max} &= \|D_{22}^{-1} W_{22} + \cdots + (D_{22}^{-1} W_{22})^l\|_{\max} \\ &\leq \|D_{22}^{-1} W_{22}\|_{\max} + \cdots + \|(D_{22}^{-1} W_{22})^l\|_{\max} \\ &\leq \frac{M}{n h_n^d} \left(1 + \cdots + (\frac{mM}{n h_n^d})^{l-1} \right) \quad \text{with probability approaching 1.} \end{aligned}$$

$$\begin{aligned} \lim_{l \rightarrow \infty} \|S_l\|_{\max} &\leq \lim_{l \rightarrow \infty} \frac{M}{n h_n^d} \left(1 + \cdots + (\frac{mM}{n h_n^d})^{l-1} \right) \\ &\leq \frac{M}{n h_n^d} / (1 - \frac{mM}{n h_n^d}) \leq \frac{2M}{n h_n^d} \quad \text{with probability approaching 1.} \end{aligned}$$

Thus $S \triangleq \lim_{l \rightarrow \infty} S_l$ exists with probability approaching 1 since $\lim_{l \rightarrow \infty} \|S_l\|_{\max} < \infty$, and S also has tiny elements. Therefore,

$$(D_{22} - W_{22})^{-1} = (I_m - D_{22}^{-1} W_{22})^{-1} D_{22}^{-1} = (I_m + S) D_{22}^{-1},$$

with probability approaching 1.

We now go back to the solution of the hard criterion of graph-based semi-supervised learning,

$$\begin{aligned} \hat{\mathbf{f}}_{(n+1):(n+m)} &= (D_{22} - W_{22})^{-1} W_{21} \mathbf{Y}_n \\ &= (I_m + S) D_{22}^{-1} W_{21} \mathbf{Y}_n = D_{22}^{-1} W_{21} \mathbf{Y}_n + S D_{22}^{-1} W_{21} \mathbf{Y}_n, \end{aligned} \quad (12)$$

with probability approaching 1. For $1 \leq a \leq m$, $\hat{f}_{(n+a)}$ equals to the a th row of $(D_{22} - W_{22})^{-1} W_{21} \mathbf{Y}_n$, i.e.,

$$\begin{aligned}\hat{f}_{(n+a)} &= \{(D_{22} - W_{22})^{-1} W_{21} \mathbf{Y}_n\}_a \\ &= \sum_{i=1}^n \frac{w_{i,n+a}}{d_{n+a,n+a}} Y_i + (S)_a D_{22}^{-1} W_{21} \mathbf{Y}_n,\end{aligned}\tag{13}$$

with probability approaching 1.

By assumption, Y_i 's are bounded. Without loss of generality, assume $\|Y_n\|_{\max} \leq 1$. For $1 \leq a \leq m$, define

$$g_{(n+a)} = \sum_{i=1}^n Y_i \left(\frac{w_{i,n+a}}{\sum_{k=1}^n w_{k,n+a}} - \frac{w_{i,n+a}}{d_{n+a,n+a}} \right).$$

We have

$$\begin{aligned}0 \leq g_{(n+a)} &\leq \sum_{i=1}^n \|Y_n\|_{\max} \left(\frac{w_{i,n+a}}{\sum_{k=1}^n w_{k,n+a}} - \frac{w_{i,n+a}}{d_{n+a,n+a}} \right) \\ &= \frac{\sum_{i=1}^n w_{i,n+a}}{\sum_{k=1}^n w_{k,n+a}} - \frac{\sum_{i=1}^n w_{i,n+a}}{\sum_{k=1}^{n+m} w_{k,n+a}} \\ &= \frac{\sum_{k=n+1}^{n+m} w_{k,n+a}}{d_{n+a,n+a}} \\ &\leq \frac{mk}{\beta \Phi_n(a) n p_n} \leq \frac{mM}{nh_n^d} \rightarrow 0,\end{aligned}$$

with probability approaching 1 as $n \rightarrow \infty$. This implies

$$g_{(n+a)} \rightarrow 0 \text{ in probability,}$$

since for any $\epsilon > 0$ we can find $m, n \in \mathbb{N}$ such that $\frac{mM}{nh_n^d} \leq \epsilon$ and

$$\mathbb{P}(|g_{(n+a)}| \leq \epsilon) \geq \mathbb{P}\left(|g_{(n+a)}| \leq \frac{mM}{nh_n^d}\right) \rightarrow 1.$$

Finally, for each $1 \leq a \leq m$,

$$\begin{aligned}\hat{f}_{(n+a)} &= \sum_{i=1}^n \frac{w_{i,n+a}}{d_{n+a,n+a}} Y_i + (S)_a D_{22}^{-1} W_{21} \mathbf{Y}_n \\ &= \sum_{i=1}^n \frac{w_{i,n+a}}{\sum_{k=1}^n w_{k,n+a}} Y_i + (S)_a D_{22}^{-1} W_{21} \mathbf{Y}_n - g_{(n+a)}.\end{aligned}$$

Since S has tiny elements,

$$\|(S)_a D_{22}^{-1} W_{21} \mathbf{Y}_n\|_{\max} \leq \frac{mM}{nh_n^d} \rightarrow 0 \text{ with probability approaching 1,}$$

which implies $(S)_a D_{22}^{-1} W_{21} \mathbf{Y}_n \rightarrow 0$ in probability. The theorem then holds by the consistency of Nadaraya-Watson estimator.

4 Numerical Studies

In this section, we compare the performance of the hard criterion and the soft criterion with different tuning parameters under a linear and non-linear model.

The inputs X_1, \dots, X_{n+m} are generated independently from a truncated multivariate normal distribution. Specifically, let \tilde{X}_i follow a p -dimensional multivariate normal with the mean $\mu = (0.5, \dots, 0.5)$, and the variance-covariance matrix

$$\begin{pmatrix} 0.1 & 0.05 & 0.05 & \dots & 0.05 \\ 0.05 & 0.1 & 0.05 & \dots & 0.05 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.05 & 0.05 & 0.05 & \dots & 0.1 \end{pmatrix}.$$

We set $p = 5$. For $i = 1, \dots, n + m$ and $k = 1, \dots, p$, let $X_{ik} = \tilde{X}_{ik}$ if $\tilde{X}_{ik} \in [0, 1]$ and $\tilde{X}_{ik} = 0$ otherwise, where X_{ik} and \tilde{X}_{ik} are the k -th component of X_i and \tilde{X}_i , respectively.

Let W be the Gaussian radial basis function (RBF) kernel, that is,

$$w_{ij} = \exp\left(-\frac{\|X_i - X_j\|^2}{\sigma^2}\right), \text{ for } 1 \leq i, j \leq m + n,$$

where $\sigma = h_n = (\log n/n)^{1/5}$. Note that W has compact support since X_i 's are truncated, and the choice of h_n satisfies the condition in Theorem 1.

We consider two models in simulation studies. In Model 1, the responses Y_i 's follow a logistic regression with

$$\text{logit } q(X_i) = -1.35 + 2X_{i1} - X_{i2} + X_{i3} - X_{i4} + 2X_{i5},$$

for $i = 1, \dots, m + n$. Model 2 uses a non-linear logit function,

$$\text{logit } q(X_i) = -1.35 + 2X_{i1} - X_{i2} + X_{i3} - X_{i4} + 2X_{i5} + X_{i1}X_{i3} + X_{i2}X_{i4},$$

for $i = 1, \dots, m + n$.

We compare the performance of graph-based learning methods with four different tuning parameters, $\lambda = 0, 0.01, 0.1$ and 5 . The performance is measured by the root mean squared error (RMSE) on the unlabeled data, that is,

$$\sqrt{\frac{1}{m} \sum_{a=1}^m (q(X_{n+a}) - \hat{q}_{n+a})^2}.$$

Each simulation is repeated 1000 times and the average RMSEs are reported.

Figure 1 shows the RMSEs under Model 1 when the sample size of unlabeled data m is fixed as 30 and the sample size of labeled data $n = 10, 30, 50, 100, 200, 300, 500, 800, 1000$ and 1500 . As n increases, the RMSEs of all methods decrease as expected. More importantly, the RMSE increases

as λ increases. In particular, the hard criterion always outperforms the soft criterion, which is consistent with our theoretical results.

Figure 2 shows the RMSEs under Model 1 when n is fixed as 100 and $m = 30, 60, 100, 300, 500$ and 1000. As before, the RMSE always increases as λ increases. Moreover, the RMSEs of all methods increase as m increases, which suggests that the hard criterion may not be consistent when m grows faster than n . For the non-linear logit function, Figure 3 and 4 show the same patterns as in Figure 1 and 2, respectively, which also support our theoretical results.

5 Summary

In this article, we proved the consistency of graph-based semi-supervised learning when the tuning parameter of the graph Laplacian is zero (the hard criterion) and showed that the method can be inconsistent when the tuning parameter is nonzero (the soft criterion). Moreover, the numerical studies also suggest that the hard criterion outperforms the soft criterion in terms of the RMSE. These results provide a better understanding about the statistical properties of graph-based semi-supervised learning. Of course, the accuracy of prediction can be measured by other indicators such as the area under the receiver operating characteristic curve (AUC). The hard criterion may not always be the best choice in term of these indicators. Further theoretical properties such as rank consistency will be explored in future research. Moreover, we would also like to investigate the behavior of these methods when the unlabeled data grow faster than the label data.

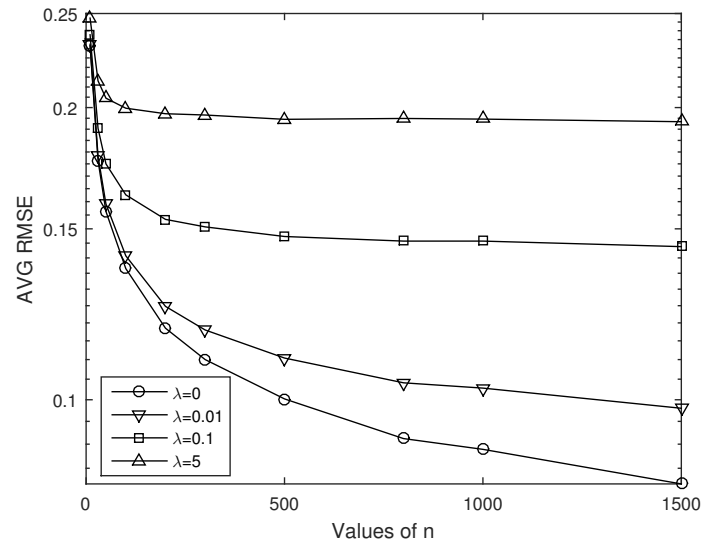


Figure 1: Average RMSEs when $m = 30$ under Model 1

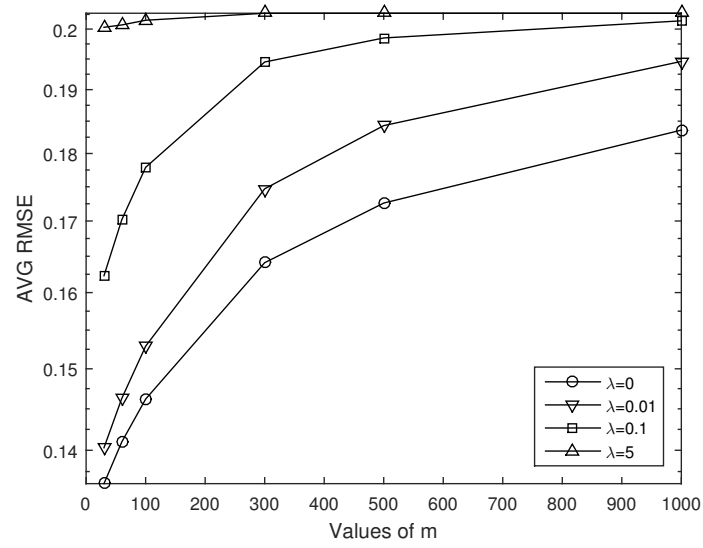


Figure 2: Average RMSEs when $n = 100$ under Model 1

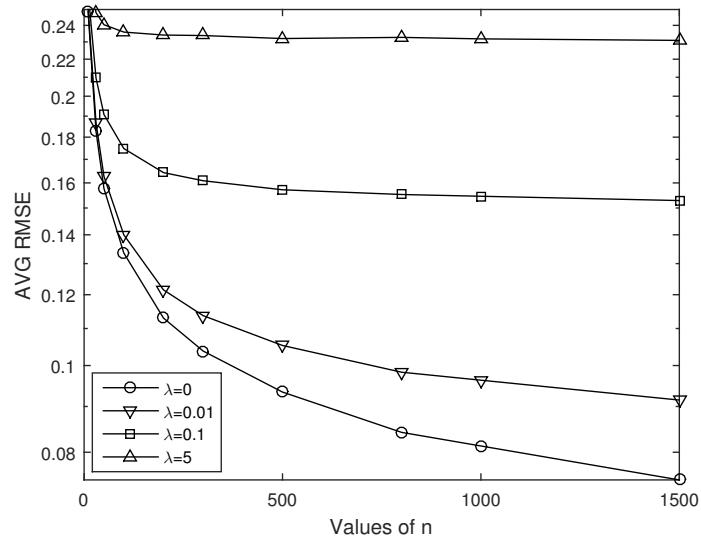


Figure 3: Average RMSEs when $m = 30$ under Model 2

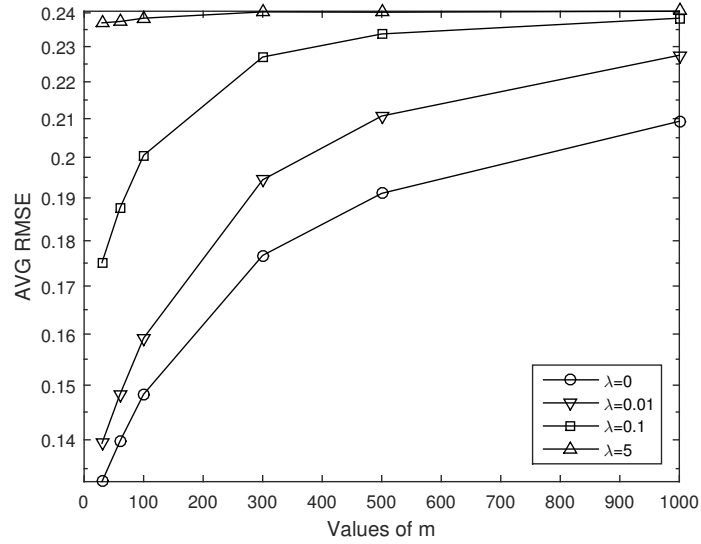


Figure 4: Average RMSEs when $n = 100$ under Model 2

References

- [1] S. Basu, A. Banerjee, and R. J. Mooney. Semi-supervised clustering by seeding. *In Proceedings of the International Conference on Machine Learning*, pages 19–26, 2002.
- [2] M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. *In Proceedings of the Seventeenth Annual Conference on Computational Learning Theory*, pages 624–638, Banff, Canada, 2004.

- [3] M. Belkin, P. Niyogi, and S. Sindhwani. Manifold Regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.
- [4] T. De Bie and N. Cristianini. Convex methods for transduction. *In Advances in Neural Information Processing Systems*, 16:73–80, 2004.
- [5] O. Bosquet, O. Chapelle, and M. Hein. Measure based regularization. *NIPS*, 16, 2004.
- [6] Zongwu Cai. Weighted Nadaraya Watson regression estimation. *Statistics & Probability Letters*, 51:307–318, 2001.
- [7] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised Learning*. The MIT Press, 2006.
- [8] Olivier Delalleau, Yoshua Bengio, and Nicolas Le Roux. Efficient non-parametric function induction in semi-supervised learning. *In Artificial Intelligence and Statistics*, 2005.
- [9] Luc. P. Devroye. The uniform convergence of the Nadaraya-Watson regression function estimate. *The Canadian Journal of Statistics*, 6:179–191, 1978.
- [10] Luc. P. Devroye and T. J. Wagner. Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, 8(2):231–239, 1980.
- [11] M. Hein. Uniform convergence of adaptive graph-based regularization. *COLT: Learning Theory*, pages 50–64, 2006.
- [12] M.D. Intriligator and Z. Griliches. *Handbook of Econometrics*, volume 1. North-Holland Publishing Company, 1988.
- [13] Rosie Jones. Learning to extract entitles from labeled and unlabeled text. *PhD Thesis*, 2005.
- [14] John Lafferty and Larry Wasserman. Statistical analysis of semi-supervised regression. *NIPS*, 20, 2008.
- [15] E. A. Nadaraya. On estimating regression. *Theor. Probability Appl*, 9:141–142, 1964.
- [16] Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Semi-supervised learning with the Graph Laplacian: The limit of infinite unlabelled data. *NIPS*, 2009.
- [17] M. E. J. Newman. *Networks: An introduction*. Oxford University Press, 2010.
- [18] Joel Ratsaby and Santosh S. Venkatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. *In Proceedings of COLT '95 Proceedings of the eighth annual conference on Computational learning theory*, pages 412–417, 1995.
- [19] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. *In Proceedings of Seventh IEEE Workshop on Applications of Computer Vision*, 2005.
- [20] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- [21] G. S. Watson. Smooth regression analysis. *Sankhyā, Series A*, 26:359–372, 1964.
- [22] D. Werner. *Functional Analysis (in German)*. Springer Verlag, 2005.

- [23] Y. Zhang, M. Brady, and S. Smith. Hidden markov random field model and segmentation of brain mr images. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.
- [24] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. in S. Thrun, L. Saul, and B. Schölkopf, editors. *MIT Press, Cambridge, MA*, 2004.
- [25] X. Zhu, Zoubin Ghahramani., and John Lafferty. Semi-supervised learning using Gaussian Fields and Harmonic Functions. *ICML*, (118), 2003.
- [26] X. Zhu and Andrew B. Goldberg. *Introduction to Semi-supervised Learning*. Morgan & Claypool Publishers, 2009.